

MULTIMEDIA



UNIVERSITY

STUDENT ID NO

--	--	--	--	--	--	--	--	--	--

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2018/2019 SESSION

TDS 3301 – DATA MINING

(All Sections / Groups)

25 OCT 2018
2:30 p.m – 4:30 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 5 printed pages including cover page with 4 questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of marks for each question is given.
3. Please write all your answers in the Answer Booklet provided

Question 1

A	B	C	D	E	F	G	H	I
Date_Start	Date_End	Country	Location	Type	Sub_Type	Killed	Cost	Disaster
102008	102008	Afghanistan	Kunduz, Balkh, Faryab, Ba ...	Drought	Drought		280000	2008-9475
72006	2006	Afghanistan		Drought	Drought		1900000	2006-9570
52000	2002	Afghanistan	Yandahar, Helmand, Nimroz	Drought	Drought	37	2580000	2000-9186
81971	1973	Afghanistan	Central, North-West, Nort ...	Drought	Drought			1971-9085
11969	1969	Afghanistan	Paktia province	Drought	Drought		48000	1969-9007
29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	Earthquake (seismic activity)	Earthquake (ground shaking)	1	935	2006-0405
13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	Earthquake (ground shaking)	5	501	2005-0586
8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	Earthquake (ground shaking)	1		2005-0575
18072004	18072004	Afghanistan	Paktia province	Earthquake (seismic activity)	Earthquake (ground shaking)	2	1040	2004-0436
10042003	10042003	Afghanistan	Yakabagh (Takhar province ...	Earthquake (seismic activity)	Earthquake (ground shaking)	1	1001	2003-0236
30112005	3122005	Albania	Vlora, Fie, Gjirokaster, ...	Flood	General flood	3	500	2005-0696
4122004	8122004	Albania	Obot (Shkodra Prefecture) ...	Flood	General flood		2500	2004-0633
21092002	10102002	Albania	Lezha, Shkoder regions (N ...	Flood	General flood	1	66884	2002-0607
20121997	23121997	Albania	Lezhe (North-Western)	Flood	Storm surge/coastal flood		8000	1997-0302
27121995	27121995	Albania	Shkadra, Malesi, Modhe, ...	Flood	General flood		2000	1995-0300
20091995	20091995	Albania	Laci, Rrogozhina, Lushnja ...	Flood		4	1500	1995-0234
17111992	19111992	Albania	Kruja, Lac, Lezha, Shkodr ...	Flood	Flash flood	11	35000	1992-0160
7121991	7121991	Albania	Fushe Arrez	Industrial Accident	Fire	60		1991-0395

- a) Your project manager is analysing a disaster data set and now he is requesting you to do as follows.
- He wants you to aggregate "Country" and "Killed". Explain how you can complete the task. Provide the partial content of the excel file to help your explanation. (2 marks)
 - He does not want redundant information. Suggest two features to be removed. (2 marks)
 - He does not want to see too many unique records for a particular piece of information. Suggest one feature to remove. (1 mark)
- b) A contingency table is as shown below.

	Student	Non-Student
Preferred sport	11	4
Preferred gaming	3	8

- Calculate the expected frequency for each cell. (2 marks)
- Calculate Chi-square statistic. Show your steps. (2 marks)
- What can you conclude from the answer in b(ii) on the null hypothesis of independence at a confidence level of 0.1? (1 mark)

Continued...

Question 2

=== Confusion Matrix ===

```
a b <-- classified as
14 4 | a = yes
 6 4 | b = no
```

The above is the confusion matrix obtained after running C4.5 algorithm on a data set. Let "yes" be positive and "no" be negative.

- What is the total number of records in the data set? (1 mark)
- How many records are labelled as "yes" in the data set? (1 mark)
- What is the number of false positive? (1 mark)
- Calculate the accuracy. Show your steps. (2 mark)
- Calculate recall and precision for class "Yes". Show your steps. (4 marks)
- What kind of relationship will you usually see between recall and precision? (1 mark)

Continued...

Question 3

- a) The following 5 records are selected from a pool of data. Calculate the true positive and false positive rates. Then draw a Receiver Operating Characteristics (ROC) curve for the following records. (2.5 + 2 marks)

Record	Class	Probability	TP	FP	TN	FN
1	P	0.80	2	0	5	3
2	P	0.60	3	1	4	2
3	N	0.54	4	2	3	1
4	N	0.51	4	4	1	1
5	N	0.40	5	5	0	0

- b) Information gain used in ID3 algorithm is biased towards multi-valued features. Suggest a method to solve the problem. (0.5 mark)
- c) A decision tree constructed for classification may over-fit a training data. What are the possible reasons of over-fitting? What is the effect of employing such a decision tree in classifying new records? (2 marks)
- d) The following data show feature values that are related to these three different species of the iris plants: iris-setosa, iris-versicolor and iris-virginica.

sepalength	sepalwidth	petallength	petalwidth	class
inf_5.4	Same	inf-2.75	inf-0.75	Iris-setosa
inf_5.4	Same	inf-2.75	inf-0.75	Iris-setosa
inf_5.4	Same	inf-2.75	inf-0.75	Iris-setosa
inf_5.4	Same	inf-2.75	inf-0.75	Iris-setosa
5.4-6	Same	2.75-4.7	0.75-inf	Iris-versicolor
5.4-6	Same	2.75-4.7	0.75-inf	Iris-versicolor
5.4-6	Same	2.75-4.7	0.75-inf	Iris-versicolor
6-inf	Same	4.7-inf	0.75-inf	Iris-virginica
6-inf	Same	4.7-inf	0.75-inf	Iris-virginica
6-inf	Same	4.7-inf	0.75-inf	Iris-virginica

- i) Suggest an unnecessary feature that can be removed from the data? Justify your choice. (0.5+0.5 mark)
- ii) Analyse the data carefully and draw a decision tree to classify these iris plants. (2 marks)

Continued...

Question 4

- Explain in brief an example of an application that uses clustering in business domains. (2 marks)
- Can clustering be used for pre-processing in data mining? Explain your answer. (2 marks)
- Use k-means algorithm with Euclidean distance, cluster the following data into two groups. The randomly selected centroids are record number 2 and 6. Show only the new centroids after the first round of execution. Show your steps. (5 marks)

Record	X	Y
1	12	25
2	29	40
3	29	50
4	29	60
5	29	70
6	26	80

- Why is k-means algorithm sensitive to noisy data and outliers? (1 mark)

Formulae:

Expected frequencies (E), (Column total x Row total)/N

Chi-square statistic, $\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$

Accuracy = (TP+TN)/ALL

Recall = TP/(TP+FN)

Precision = TP/(TP+FP)

Euclidean distance = $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Degrees of Freedom	Chi-Square (χ^2) Distribution Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209

End of Page